

DISCRETIZATION OF LINEAR PROBLEMS IN BANACH SPACES: RESIDUAL MINIMIZATION, NONLINEAR PETROV–GALERKIN, AND MONOTONE MIXED METHODS

I. Muga^{*} and K.G. van der Zee[†]

— 16th November, 2015 —

Abstract

This work presents a comprehensive optimal-discretization theory for linear equations in Banach spaces. As part of our theory, a class of nonlinear Petrov–Galerkin projectors is studied, which are key in establishing optimal a priori error estimates involving constants depending on the geometry of the underlying Banach spaces.

Contents

1	Introduction	2
2	Linear equations in Banach spaces	5
2.1	Statement of problem	5
2.2	Abstract operator equations in Banach spaces	5
2.3	Analysis of problem	6
3	Residual minimization in Banach spaces	7
3.1	Statement of discrete problem	7
3.2	Best approximation in Banach spaces	8
3.3	Analysis of residual minimization	8

^{*}Pontificia Universidad Católica de Valparaíso, Instituto de Matemáticas, ignacio.muga@uvc.cl

[†]University of Nottingham, School of Mathematical Sciences, kg.vanderzee@nottingham.ac.uk

4	Theory of duality mappings	10
4.1	The duality mapping	10
4.2	Special Banach spaces	11
4.3	Subdifferential characterization and minimization	13
5	Nonlinear Petrov–Galerkin method and monotone mixed formulation	15
5.1	Characterization of residual minimization	16
5.2	The reflexive smooth setting	16
6	The inexact monotone mixed method	18
6.1	Equivalent discrete settings	19
6.2	Well-posedness of the inexact method	20
6.3	Error analysis of the inexact method	22
6.4	Direct a priori error analysis of the inexact method	23
	Acknowledgements	27

1 Introduction

In the setting of *Banach* spaces, we consider the abstract problem

$$\text{Find } u \in \mathbb{U} \text{ such that } Bu = f \quad \text{in } \mathbb{V}^*, \quad (1)$$

where \mathbb{U} and \mathbb{V} are Banach spaces, $B : \mathbb{U} \rightarrow \mathbb{V}^*$ is a continuous, bounded-below, linear operator, and the data f is a given element in the dual space \mathbb{V}^* . For a given discrete subspace $\mathbb{U}_n \subset \mathbb{U}$, of dimension n , the objective of this paper is to present a Galerkin-based discretization technique which is guaranteed to provide a near-best approximation $u_n \in \mathbb{U}_n$ to the solution u , i.e., u_n satisfies the a priori error estimate

$$\|u - u_n\|_{\mathbb{U}} \leq C \inf_{w_n \in \mathbb{U}_n} \|u - w_n\|_{\mathbb{U}},$$

for some constant $C \geq 1$, independent of n , in which case the discretization method is said to be *(quasi-) optimal*.

The fundamental difficulty in the design of a discretization technique for the problem in (1) is *stability*. For example, in a *standard* Petrov–Galerkin discretization, where $u_n \in \mathbb{U}_n$ satisfies

$$\langle Bu_n - f, v_n \rangle_{\mathbb{V}^*, \mathbb{V}} = 0 \quad \forall v_n \in \mathbb{V}_n, \quad (2)$$

it is well-known that one must come up with a test space $\mathbb{V}_n \subset \mathbb{V}$ that is *compatible* with \mathbb{U}_n in the sense that the discrete inf-sup conditions are satisfied; see, e.g., [12, Lemma 2.28].

When \mathbb{U} and \mathbb{V} are *Hilbert* spaces, a theory of optimal Petrov–Galerkin discretizations has recently been erected in a pioneering sequence of papers by Demkowicz, Gopalakrishnan, and others; see, e.g., [10] and the recent overview in [11]. A main result within that theory states that the Petrov–Galerkin method (2) with the (intractable) test space

$$\mathbb{V}_n = R_{\mathbb{V}}^{-1} B \mathbb{U}_n, \quad (3)$$

where $R_{\mathbb{V}} : \mathbb{V} \rightarrow \mathbb{V}^*$ is the Riesz map, *guarantees* a stable method yielding a near-best approximation u_n . In fact, in that case, u_n is the best approximation when measured in $\|B(\cdot)\|_{\mathbb{V}^*}$, which is a norm on \mathbb{U} that can be shown to be equivalent to $\|\cdot\|_{\mathbb{U}}$.

Various interpretations and connections to other discretization techniques are possible for the above optimal Petrov–Galerkin method (2)–(3) in Hilbert spaces, a summary of which can be found in [11]. Let us briefly mention those that are relevant to our study. The optimal Petrov–Galerkin method can be interpreted as a residual-minimization method in the dual space \mathbb{V}^* (minimizing $w_n \mapsto \|f - Bw_n\|_{\mathbb{V}^*}$) as clarified first in [10]. It can also be equivalently formulated as a (semi-infinite) saddle-point problem on $\mathbb{V} \times \mathbb{U}_n$ which includes $r = R_{\mathbb{V}}^{-1}(f - Bu_n) \in \mathbb{V}$ as an auxiliary unknown, a connection which was discovered by Cohen, Dahmen and Welper [8].

Depending on the viewpoint, different techniques are possible for the approximation of the intractable \mathbb{V}_n in (3) giving rise to *inexact* methods. The DPG (discontinuous Petrov–Galerkin) finite element method approximates (2) by projecting element-wise the test functions in \mathbb{V}_n onto enriched subspaces, an operation which is computationally feasible by virtue of a hybrid formulation where \mathbb{V} is a broken Sobolev space. Alternatively, an inexact form of the saddle-point problem on $\mathbb{V} \times \mathbb{U}_n$ is obtained by replacing \mathbb{V} by a discrete subspace \mathbb{V}_m leading to a fully-discrete mixed method. Sufficient conditions for the stability and (quasi-) optimality of inexact methods can be found in [14, 13, 4].

The current optimal theory is a *Hilbert* Petrov–Galerkin theory, for which the setting is based on Hilbert spaces. Indeed, the theory crucially relies on the Riesz map in Hilbert spaces, and, therefore, it does not directly apply to the more general setting of Banach spaces. Such a more general setting is necessary for example for PDEs posed in non-Hilbert Sobolev spaces (e.g., to handle rough data or nonlinearities), and those problems would benefit from an optimal *Banach* Petrov–Galerkin theory.

In this work, we consider the extension of optimal Petrov–Galerkin discretization methods from Hilbert spaces to Banach spaces. Indispensable in our construction is the (multi-valued) *duality mapping* $\mathcal{J}_{\mathbb{V}} : \mathbb{V} \rightarrow 2^{\mathbb{V}^*}$, which is the operator that extends the Riesz map. Fundamentally, our theory corresponds to that of residual minimization, a concept which does not rely on any Hilbert-space construct. Therefore, theoretically,

$$u_n = \arg \min_{w_n \in \mathbb{U}_n} \|f - Bw_n\|_{\mathbb{V}^*}.$$

Originally, this idea goes back to Guermond [15], who considered residual minimization in $\mathbb{V}^* = L^p(\Omega)$, for $1 \leq p < \infty$.

As the first main result in this paper, we shall prove fundamental characterizations of the residual minimizer in abstract settings employing the duality mapping. In particular, in the case of reflexive smooth Banach spaces, for which the duality mapping is single-valued and denoted by $J_{\mathbb{V}} : \mathbb{V} \rightarrow \mathbb{V}^*$, we establish the equivalence with the optimal Petrov–Galerkin discretization:

$$\langle J_{\mathbb{V}}^{-1}(f - Bu_n), \nu_n \rangle_{\mathbb{V}, \mathbb{V}^*} = 0 \quad \forall \nu_n \in B\mathbb{U}_n,$$

which is equivalent to

$$\langle J_{\mathbb{V}}^{-1}(f - Bu_n), Bw_n \rangle_{\mathbb{V}, \mathbb{V}^*} = 0 \quad \forall w_n \in \mathbb{U}_n.$$

Since $J_{\mathbb{V}}$ is nonlinear unless \mathbb{V} is a Hilbert space, this represents a *nonlinear* Petrov–Galerkin method, which is an extension of (2)–(3) to Banach spaces. Furthermore, analogous to the Hilbert-space case, we shall establish the equivalence with a (semi-infinite) saddle-point problem on $\mathbb{V} \times \mathbb{U}_n$. Because of nonlinearity, however, the associated mixed formulation is now *monotone*.

The second main result in this paper is the analysis of a *tractable* inexact method based on the monotone mixed formulation. We will demonstrate discrete equivalences, provide conditions guaranteeing the existence of discrete solutions, and prove optimal a priori and a posteriori error estimates. The analysis implies that the inexact method is stable and quasi-optimal, thereby providing near-best approximations. In our analysis we crucially rely on a technique involving the Fortin operator, similar to the analyses in [14, 4], and a recent projection result by Stern [20]. Moreover, to obtain a sharper a priori error estimate, we proof a novel a priori bound for nonlinear Petrov–Galerkin projectors. This a priori bound is of independent interest and applies to general best-approximation minimizers.

The paper is outlined as follows. We introduce our problem of interest in Section 2 and recall elementary results for its well-posedness. Next, we consider residual minimization in Section 3. To be able to analyze residual minimization,

we recall essential theory of duality mappings in Section 4, and we prove a sharpened a priori bound for best-approximation minimizers. Section 5 then considers the corresponding nonlinear Petrov–Galerkin method and monotone mixed formulation. Finally, in Section 6, we analyze the inexact monotone mixed method.

We employ standard notation throughout this paper. If \mathbb{X} is any Banach space, the norm of \mathbb{X} will be denoted by $\|\cdot\|_{\mathbb{X}}$, the dual space by \mathbb{X}^* , and the dual norm by $\|\cdot\|_{\mathbb{X}^*}$. The duality pairing between \mathbb{X}^* and \mathbb{X} will be denoted by $\langle \cdot, \cdot \rangle_{\mathbb{X}^*, \mathbb{X}}$. For a subspace $\mathbb{M} \subset \mathbb{X}$, the orthogonal space to \mathbb{M} is defined by $\mathbb{M}^\perp := \{x^* \in \mathbb{X}^* : \langle x^*, x \rangle_{\mathbb{X}^*, \mathbb{X}} = 0, \forall x \in \mathbb{M}\} \subset \mathbb{X}^*$. However, for a subspace of a dual space, $\mathbb{N} \subset \mathbb{X}^*$, the orthogonal space to \mathbb{N} is defined by $\mathbb{N}^\perp := \{x \in \mathbb{X} : \langle x^*, x \rangle = 0, \forall x^* \in \mathbb{N}\} \subset \mathbb{X}$.

2 Linear equations in Banach spaces

In this section we introduce our problem of interest. Motivated by weak formulations of PDEs, we start with a variational form.

2.1 Statement of problem

Let \mathbb{U} and \mathbb{V} be two Banach spaces and $b : \mathbb{U} \times \mathbb{V} \rightarrow \mathbb{R}$ a continuous bilinear form. Given any $f \in \mathbb{V}^*$, the variational form of our interest is:

$$\text{Find } u \in \mathbb{U} \text{ such that } b(u, v) = \langle f, v \rangle_{\mathbb{V}^*, \mathbb{V}} \quad \forall v \in \mathbb{V}. \quad (4)$$

The continuous bilinear form b defines a bounded linear operator $B : \mathbb{U} \rightarrow \mathbb{V}^*$ by

$$\langle Bw, v \rangle_{\mathbb{V}^*, \mathbb{V}} := b(w, v), \quad \forall v \in \mathbb{V}, w \in \mathbb{U}. \quad (5)$$

Therefore (4) is equivalent to:

$$\text{Find } u \in \mathbb{U} \text{ such that } Bu = f \text{ in } \mathbb{V}^*. \quad (6)$$

2.2 Abstract operator equations in Banach spaces

To analyze (6), we briefly recall elementary theory for an abstract linear problem.

Let \mathbb{X} and \mathbb{Y} be two Banach spaces and $A : \mathbb{X} \rightarrow \mathbb{Y}$ a linear operator. Given any $y \in \mathbb{Y}$, a general linear problem in Banach spaces is:

$$\text{Find } x \in \mathbb{X} \text{ such that } Ax = y. \quad (7)$$

If there is a constant $C > 0$ such that $\|Ax\|_{\mathbb{Y}} \leq C\|x\|_{\mathbb{X}}$ for all $x \in \mathbb{X}$, then the operator A is *bounded* (or *continuous*), and M_A , defined as the smallest possible constant C , is called the continuity constant of A . If there is a constant $C > 0$

such that $\|Ax\|_{\mathbb{Y}} \geq C\|x\|_{\mathbb{X}}$ for all $x \in \mathbb{X}$, then A is *bounded below*. In that case, the largest possible constant C will be called the *bounded-below constant* and denoted by γ_A . The range of A is denoted by $\text{Im}(A) := A(\mathbb{X}) \subseteq \mathbb{Y}$. The *adjoint* of A is $A^* : \mathbb{Y}^* \rightarrow \mathbb{X}^*$. The kernel of A^* , denoted $\text{Ker}(A^*)$, consists of all $y^* \in \mathbb{Y}^*$ such that $A^*y^* = 0$.

Proposition 2.1 *Let $A : \mathbb{X} \rightarrow \mathbb{Y}$ be a linear, continuous and bounded below operator.*

- (i) *If $y \in \text{Im}(A)$, then there exists a unique solution $x \in \mathbb{X}$ to problem (7), which moreover satisfies the a priori estimate*

$$\|x\|_{\mathbb{X}} \leq \frac{1}{\gamma_A} \|y\|_{\mathbb{Y}}. \quad (8)$$

- (ii) *A is surjective if and only if $\text{Ker}(A^*) = \{0\}$, in which case the result at (i) applies to any $y \in \mathbb{Y}$.* □

Proof The proof is classical; see, e.g., Ern and Guermond [12, Appendix A.2] or Oden and Demkowicz [16, Sec. 5.17]. ■

2.3 Analysis of problem

We now return to our original problem in (4) by applying Proposition 2.1(i) to its operator form in (6).

Proposition 2.2 *Let $b : \mathbb{U} \times \mathbb{V} \rightarrow \mathbb{R}$ be a continuous bilinear form such that¹*

$$\gamma_b := \inf_{w \in \mathbb{U}} \sup_{v \in \mathbb{V}} \frac{b(w, v)}{\|w\|_{\mathbb{U}} \|v\|_{\mathbb{V}}} > 0. \quad (9a)$$

If $f \in \text{Im}(B)$, with B defined by (5), then there exists a unique solution $u \in \mathbb{U}$ to (4) (or, equivalently, (6)), which moreover satisfies the a priori estimate:

$$\|u\|_{\mathbb{U}} \leq \frac{1}{\gamma_b} \|f\|_{\mathbb{V}^*}. \quad \square$$

Proof Condition (9a) implies that B is bounded below with constant $\gamma_B = \gamma_b$. We then apply Proposition 2.1(i). ■

¹In this paper we systematically abuse the inf and sup notation for fractions by writing, e.g., $\sup_{v \in \mathbb{V}} \frac{b(w, v)}{\|v\|_{\mathbb{V}}}$ instead of $\sup_{v \in \mathbb{V} \setminus \{0\}} \frac{b(w, v)}{\|v\|_{\mathbb{V}}}$.

Remark 2.3 By Proposition 2.1(ii), B is surjective if and only if $B^* : \mathbb{V}^{**} \rightarrow \mathbb{U}^*$ is injective, in which case Proposition 2.2 holds for every $f \in \mathbb{V}^*$. If \mathbb{V} is reflexive, then B^* coincides with the *dual operator* $B^* : \mathbb{V} \rightarrow \mathbb{U}^*$ defined by

$$\langle B^*v, w \rangle_{\mathbb{U}^*, \mathbb{U}} := b(w, v), \quad \forall w \in \mathbb{U}, v \in \mathbb{V}. \quad \square$$

Injectivity of B^* is then equivalent to the following condition on b :

$$\left\{ v \in \mathbb{V} : b(w, v) = 0, \forall w \in \mathbb{U} \right\} = \{0\}. \quad (9b)$$

Remark 2.4 The constant γ_b in Proposition 2.2 is commonly referred to as the inf-sup constant. Eqs. (9a) and (9b) are commonly referred to as the inf-sup conditions (when \mathbb{V} is reflexive) [12]. \square

3 Residual minimization in Banach spaces

As the fundamental approximation to problem (4), we consider the minimization of the residual of the equation in (6). In this section we analyze this discretization method, in particular, we show that the method is stable and quasi-optimal.

3.1 Statement of discrete problem

Given a finite-dimensional subspace $\mathbb{U}_n \subset \mathbb{U}$, we define the approximation u_n by:

$$\text{Find } u_n \in \mathbb{U}_n \text{ such that } u_n = \arg \min_{w_n \in \mathbb{U}_n} \|f - Bw_n\|_{\mathbb{V}^*}. \quad (10)$$

In some applications the norm

$$\|\cdot\|_{\mathbb{E}} := \|B(\cdot)\|_{\mathbb{V}^*}$$

is referred to as the *energy norm* on \mathbb{U} . Assuming B is continuous and bounded below, the energy norm $\|\cdot\|_{\mathbb{E}}$ is equivalent to the norm $\|\cdot\|_{\mathbb{U}}$ in \mathbb{U} , i.e.,

$$\gamma_B \|w\|_{\mathbb{U}} \leq \|w\|_{\mathbb{E}} \leq M_B \|w\|_{\mathbb{U}}, \quad \forall w \in \mathbb{U}. \quad (11)$$

Therefore, if $f \in \text{Im } B$, it is clear that upon substituting $f = Bu$, (10) is equivalent to finding the best approximation to u in the energy norm:

$$\text{Find } u_n \in \mathbb{U}_n \text{ such that } u_n = \arg \min_{w_n \in \mathbb{U}_n} \|u - w_n\|_{\mathbb{E}}. \quad (12)$$

3.2 Best approximation in Banach spaces

To analyze (12), we briefly recall some elementary results from best approximation theory in Banach spaces. To address uniqueness, the following geometric condition on Banach spaces is useful.

Definition 3.1 (Strictly-convex Banach space) A Banach space \mathbb{Y} is said to be *strictly convex* if, for all $y_1, y_2 \in \mathbb{Y}$ such that $y_1 \neq y_2$ and $\|y_1\|_{\mathbb{Y}} = \|y_2\|_{\mathbb{Y}} = 1$, it holds that

$$\|\theta y_1 + (1 - \theta)y_2\|_{\mathbb{Y}} < 1, \quad \forall \theta \in (0, 1). \quad \square$$

Lemma 3.2 (Best approximation) Let \mathbb{Y} be a Banach space and let $\mathbb{M} \subset \mathbb{Y}$ be a finite-dimensional subspace. For any $y \in \mathbb{Y}$, there exists a minimizer $y_0 \in \mathbb{M}$ such that

$$y_0 = \arg \min_{z_0 \in \mathbb{M}} \|y - z_0\|_{\mathbb{Y}}. \quad (13)$$

Moreover, such a minimizer satisfies the a priori estimate $\|y_0\|_{\mathbb{Y}} \leq 2\|y\|_{\mathbb{Y}}$. Additionally, if the Banach space \mathbb{Y} is strictly convex, then the minimizer is unique. \square

Proof See, e.g., Stakgold and Holst [19, Sec. 10.2]. \blacksquare

3.3 Analysis of residual minimization

We now return to residual minimization (10), and apply Lemma 3.2 to the equivalent best-approximation problem (12).

Theorem 3.A Let \mathbb{U} and \mathbb{V} be two Banach spaces and let $B : \mathbb{U} \rightarrow \mathbb{V}^*$ be a linear, continuous and bounded-below operator with continuity constant $M_B > 0$ and bounded-below constant $\gamma_B > 0$. Given $f \in \mathbb{V}^*$ and a finite-dimensional subspace $\mathbb{U}_n \subset \mathbb{U}$, the following statements hold:

- (i) There exists a minimizer $u_n \in \mathbb{U}_n$ of problem (10).
- (ii) Any minimizer u_n of (10) satisfies the a priori estimate

$$\|u_n\|_{\mathbb{U}} \leq \frac{2}{\gamma_B} \|f\|_{\mathbb{V}^*}. \quad (14)$$

- (iii) If \mathbb{V}^* is a strictly-convex Banach space, then the minimizer u_n of (10) is unique.

(iv) If $f \in \text{Im}(B)$ and $u \in \mathbb{U}$ is the solution of the continuous problem $Bu = f$, then we have the a posteriori and a priori error estimates:

$$\|u - u_n\|_{\mathbb{U}} \leq \frac{1}{\gamma_B} \|f - Bu_n\|_{\mathbb{V}^*} \leq \frac{M_B}{\gamma_B} \inf_{w_n \in \mathbb{U}_n} \|u - w_n\|_{\mathbb{U}}. \quad (15)$$

Proof The proof can be found in Guermond [15], but we present an alternative based on Lemma 3.2.

We first consider the case that $f \in \text{Im}(B)$, in which case $Bu = f$. Since \mathbb{U} endowed with the energy norm is a Banach space, the first statement is a direct application of Lemma 3.2 by using the energy norm topology in \mathbb{U} and the equivalence between (10) and (12).

The estimate provided by Lemma 3.2 and the norm equivalence (11) show that

$$\|u_n\|_{\mathbb{U}} \leq \frac{1}{\gamma_B} \|u_n\|_{\mathbb{E}} \leq \frac{2}{\gamma_B} \|u\|_{\mathbb{E}} = \frac{2}{\gamma_B} \|f\|_{\mathbb{V}^*},$$

which proves the second statement.

If \mathbb{V}^* is strictly convex, then \mathbb{U} endowed with the energy norm is also strictly convex. Hence, by Lemma 3.2 the minimizer $u_n \in \mathbb{U}_n$ is unique, which proves the third statement.

Finally, using the norm equivalence (11), together with the minimizing property of u_n in the energy norm, we get

$$\|u - u_n\|_{\mathbb{U}} \leq \frac{1}{\gamma_B} \|u - u_n\|_{\mathbb{E}} \leq \frac{1}{\gamma_B} \|u - w_n\|_{\mathbb{E}} \leq \frac{M_B}{\gamma_B} \|u - w_n\|_{\mathbb{U}},$$

for all $w_n \in \mathbb{U}_n$, which proves the last statement.

The proof for the case that $f \notin \text{Im}(B)$ follows similarly by considering the best approximation of f in the space $B\mathbb{U}_n$. ■

Remark 3.3 Examples of non-uniqueness of the minimizer and sharpness of (14) can be constructed when the involved Banach space is not strictly convex. For example, in \mathbb{R}^2 with the norm $\|(x_1, x_2)\|_1 = |x_1| + |x_2|$, the best approximation of the point $(1, 0)$ over the line $\{(t, t) : t \in \mathbb{R}\}$ is the whole segment $\{(t, t) : t \in [0, 1]\}$. Moreover, the point $(1, 1)$ is a best approximation and $\|(1, 1)\|_1 = 2 = 2\|(0, 1)\|_1$. □

Remark 3.4 In the context of finite elements, there is a sequence $\{\mathbb{U}_h\}_{h>0}$ of finite-dimensional subspaces, $\mathbb{U}_h \subset \mathbb{U}$, and having the interpolation property

$$\inf_{w_h \in \mathbb{U}_h} \|w - w_h\|_{\mathbb{U}} \leq \varepsilon(h) \|w\|_{\mathbb{Z}}, \quad \forall w \in \mathbb{Z},$$

where $\mathbb{Z} \subset \mathbb{U}$ is a more regular subspace and $\varepsilon(h)$ is a function that is continuous at zero and $\varepsilon(0) = 0$.² This last statement, together with (15), gives a guarantee that minimizers $u_n \in \mathbb{U}_n \equiv \mathbb{U}_h$ of equation (10) converge to $u = B^{-1}f$ upon $h \rightarrow 0^+$. In other words, this is an example of the dictum: approximability and stability imply convergence. \square

Remark 3.5 As proposed in [22], if B is bijective and \mathbb{V} reflexive (hence $B^* : \mathbb{V} \rightarrow \mathbb{U}^*$ is bijective), one can endow the space \mathbb{V} with the equivalent norm

$$\|\cdot\|_{\mathbb{V}_{\text{opt}}} = \|B^*(\cdot)\|_{\mathbb{U}^*}.$$

Then, the residual minimization in $(\mathbb{V}_{\text{opt}})^*$ reduces essentially to best approximation of u in \mathbb{U} . In particular, instead of (15), one then obtains

$$\|u - u_n\|_{\mathbb{U}} = \|f - Bu_n\|_{(\mathbb{V}_{\text{opt}})^*} = \inf_{w_n \in \mathbb{U}_n} \|u - w_n\|_{\mathbb{U}}. \quad \square$$

4 Theory of duality mappings

The concept of the duality mapping is needed to characterize the solution of residual minimization. In this section we briefly review relevant theory of duality mappings; see for more details, e.g., [9, 21, 7, 6, 3]. At the end of this Section, we prove a novel a priori bound for best-approximation minimizers. In special Banach spaces this bound sharpens the a priori estimate in Lemma 3.2.

4.1 The duality mapping

Definition 4.1 Let \mathbb{Y} be a normed vector space. The multivalued mapping $\mathcal{J}_{\mathbb{Y}} : \mathbb{Y} \rightarrow 2^{\mathbb{Y}^*}$ defined by

$$\mathcal{J}_{\mathbb{Y}}(y) := \left\{ y^* \in \mathbb{Y}^* : \langle y^*, y \rangle_{\mathbb{Y}^*, \mathbb{Y}} = \|y\|_{\mathbb{Y}}^2 = \|y^*\|_{\mathbb{Y}^*}^2 \right\},$$

is the *duality mapping* on \mathbb{Y} . \square

By the Hahn-Banach extension Theorem (see, e.g., [3, Corollary 1.3]), the set $\mathcal{J}_{\mathbb{Y}}(y) \subset \mathbb{Y}^*$ is non-empty for every $y \in \mathbb{Y}$. Observe that the definition of the duality mapping depends on the norm that we use in the Banach space \mathbb{Y} . Some basic properties of $\mathcal{J}_{\mathbb{Y}}$ are summarized in the following.

Proposition 4.2 Let \mathbb{Y} be a normed vector space and $y \in \mathbb{Y}$.

²As an example, for the Sobolev space $W^{m,p}$ and piecewise polynomial finite-dimensional approximations, we refer the reader to, e.g., [2].

- (i) The set $\mathcal{J}_{\mathbb{Y}}(y) \subset \mathbb{Y}^*$ is bounded, convex, and closed.
- (ii) The duality mapping $\mathcal{J}_{\mathbb{Y}}$ is homogeneous, and it is monotone in the sense that:

$$\langle y^* - z^*, y - z \rangle_{\mathbb{Y}^*, \mathbb{Y}} \geq (\|y\|_{\mathbb{Y}} - \|z\|_{\mathbb{Y}})^2 \geq 0,$$

for all $y, z \in \mathbb{Y}$, for all $y^* \in \mathcal{J}_{\mathbb{Y}}(y)$ and for all $z^* \in \mathcal{J}_{\mathbb{Y}}(z)$.

- (iii) For any $y^* \in \mathcal{J}_{\mathbb{Y}}(y)$, its norm supremum is achieved by y , i.e.,

$$\sup_{z \in \mathbb{Y}} \frac{\langle y^*, z \rangle_{\mathbb{Y}^*, \mathbb{Y}}}{\|z\|_{\mathbb{Y}}} = \frac{\langle y^*, y \rangle_{\mathbb{Y}^*, \mathbb{Y}}}{\|y\|_{\mathbb{Y}}}. \quad (16)$$

Proof These results are classical; see, e.g., [3, Ch. 1]. ■

4.2 Special Banach spaces

We next list important properties of the duality mapping $\mathcal{J}_{\mathbb{Y}} : \mathbb{Y} \rightarrow 2^{\mathbb{Y}^*}$ in special Banach spaces.

4.2.1 Strict convexity of \mathbb{Y}^*

If \mathbb{Y}^* is strictly convex (recall Definition 3.1), then $\mathcal{J}_{\mathbb{Y}} : \mathbb{Y} \rightarrow 2^{\mathbb{Y}^*}$ is a *single-valued map*, and in that case we will use the notation:

$$J_{\mathbb{Y}} : \mathbb{Y} \rightarrow \mathbb{Y}^*, \quad \text{in other words,} \quad \mathcal{J}_{\mathbb{Y}}(y) = \{J_{\mathbb{Y}}(y)\}.$$

To see that $\mathcal{J}_{\mathbb{Y}}(y)$ is single-valued, let $\mathcal{J}_{\mathbb{Y}}(y)$ contain two different elements, say y_1^* and y_2^* . Then, since $\mathcal{J}_{\mathbb{Y}}(y)$ is a convex set (see Prop. 4.2(i)), $\theta y_1^* + (1 - \theta)y_2^* \in \mathcal{J}_{\mathbb{Y}}(y)$ for any $\theta \in (0, 1)$. Therefore, by strict convexity, we get

$$\|y\|_{\mathbb{Y}} = \|\theta y_1^* + (1 - \theta)y_2^*\|_{\mathbb{Y}^*} < \|y_1^*\|_{\mathbb{Y}^*} = \|y\|_{\mathbb{Y}},$$

which is a contradiction. Strict convexity of \mathbb{Y}^* is also a necessary condition for single-valuedness; see [9, Proposition 12.3].

Furthermore, if \mathbb{Y}^* is strictly convex, then $\mathcal{J}_{\mathbb{Y}} : \mathbb{Y} \rightarrow 2^{\mathbb{Y}^*}$ is *hemi-continuous* [9, Sec. 3.12]:

$$J_{\mathbb{Y}}(y + \lambda z) \rightharpoonup J_{\mathbb{Y}}(y) \quad \text{as } \lambda \rightarrow 0^+. \quad (17)$$

Another important property is concerned with the duality map on subspaces. We state this as the following Lemma, and we include a proof since we could not find this result in the existing literature.

Lemma 4.3 *Let \mathbb{Y} be a Banach space, \mathbb{Y}^* strictly convex, and $J_{\mathbb{Y}} : \mathbb{Y} \rightarrow \mathbb{Y}^*$ denote the duality map on \mathbb{Y} . Let $\mathbb{M} \subset \mathbb{Y}$ denote a linear subspace of \mathbb{Y} , and $J_{\mathbb{M}} : \mathbb{M} \rightarrow \mathbb{M}^*$ denote the corresponding duality map on \mathbb{M} . Then,*

$$I_{\mathbb{M}}^* J_{\mathbb{Y}} \circ I_{\mathbb{M}} = J_{\mathbb{M}},$$

where $I_{\mathbb{M}} : \mathbb{M} \rightarrow \mathbb{Y}$ is the natural injection. □

Proof Let $z \in \mathbb{M}$ and consider the linear and continuous functional $J_{\mathbb{M}}(z) \in \mathbb{M}^*$. Using the Hahn–Banach extension (see [3, Corollary 1.2]), we extend this functional to an element $\widetilde{J_{\mathbb{M}}(z)} \in \mathbb{Y}^*$ such that $\|\widetilde{J_{\mathbb{M}}(z)}\|_{\mathbb{Y}^*} = \|J_{\mathbb{M}}(z)\|_{\mathbb{M}^*}$.³ Observe that the extension satisfies

$$\begin{aligned} \|\widetilde{J_{\mathbb{M}}(z)}\|_{\mathbb{Y}^*} &= \|I_{\mathbb{M}}z\|_{\mathbb{Y}} \quad \text{and} \\ \langle \widetilde{J_{\mathbb{M}}(z)}, I_{\mathbb{M}}z \rangle_{\mathbb{Y}^*, \mathbb{Y}} &= \langle J_{\mathbb{M}}(z), z \rangle_{\mathbb{M}^*, \mathbb{M}} = \|I_{\mathbb{M}}z\|_{\mathbb{Y}}^2. \end{aligned}$$

So, as a matter of fact, $\widetilde{J_{\mathbb{M}}(z)} = J_{\mathbb{Y}}(I_{\mathbb{M}}z)$. Therefore, by the extension property of $\widetilde{J_{\mathbb{M}}(z)}$ we get

$$I_{\mathbb{M}}^* J_{\mathbb{Y}}(I_{\mathbb{M}}z) = I_{\mathbb{M}}^* \widetilde{J_{\mathbb{M}}(z)} = J_{\mathbb{M}}(z). \quad \blacksquare$$

4.2.2 Strict convexity of \mathbb{Y}

If \mathbb{Y} is strictly convex, then $\mathcal{J}_{\mathbb{Y}} : \mathbb{Y} \rightarrow 2^{\mathbb{Y}^*}$ is *injective*. In fact, if y and z are two distinct points in \mathbb{Y} , then $\mathcal{J}_{\mathbb{Y}}(y) \cap \mathcal{J}_{\mathbb{Y}}(z) = \emptyset$. Indeed, if $y^* \in \mathcal{J}_{\mathbb{Y}}(y) \cap \mathcal{J}_{\mathbb{Y}}(z)$, then $\|y\|_{\mathbb{Y}} = \|z\|_{\mathbb{Y}}$ and

$$\|y\|_{\mathbb{Y}}^2 = \langle y^*, \theta y + (1 - \theta)z \rangle_{\mathbb{Y}^*, \mathbb{Y}} \leq \|y\|_{\mathbb{Y}} \|\theta y + (1 - \theta)z\|_{\mathbb{Y}} < \|y\|_{\mathbb{Y}}^2,$$

which is a contradiction.

Furthermore, if \mathbb{Y} is strictly convex, then $\mathcal{J}_{\mathbb{Y}}$ is *strictly monotone*:

$$\langle y^* - z^*, y - z \rangle_{\mathbb{Y}^*, \mathbb{Y}} > 0, \quad \text{for all } y \neq z, \text{ any } y^* \in \mathcal{J}_{\mathbb{Y}}(y) \text{ and } z^* \in \mathcal{J}_{\mathbb{Y}}(z). \quad (18)$$

It is known that the converse holds as well: Strict monotonicity of $\mathcal{J}_{\mathbb{Y}}$ implies strict convexity of \mathbb{Y} , a result due to Petryshyn [17].

³In fact, the Hahn–Banach extension is unique on account of strict convexity of \mathbb{Y}^* .

4.2.3 Reflexivity of \mathbb{Y}

If \mathbb{Y} is a reflexive Banach space, then $\mathcal{J}_{\mathbb{Y}} : \mathbb{Y} \rightarrow 2^{\mathbb{Y}^*}$ is *surjective*. This is meant in the following sense: Every $y^* \in \mathbb{Y}^*$ belongs to a set $\mathcal{J}_{\mathbb{Y}}(y)$, for some $y \in \mathbb{Y}$. Indeed, let $\mathcal{J}_{\mathbb{Y}^*} : \mathbb{Y}^* \rightarrow \mathbb{Y}^{**}$ be the duality mapping on \mathbb{Y}^* and choose some $y^{**} \in \mathcal{J}_{\mathbb{Y}^*}(y^*)$. By reflexivity, there is a $y \in \mathbb{Y}$ such that:

$$\|y\|_{\mathbb{Y}}^2 = \|y^{**}\|_{\mathbb{Y}^{**}}^2 = \|y^*\|_{\mathbb{Y}^*}^2 = \langle y^{**}, y^* \rangle_{\mathbb{Y}^{**}, \mathbb{Y}^*} = \langle y^*, y \rangle_{\mathbb{Y}^*, \mathbb{Y}} \quad .$$

Hence $y^* \in \mathcal{J}_{\mathbb{Y}}(y)$.

4.2.4 Reflexive smooth setting

An important case in our study is when the Banach space \mathbb{Y} has all the previously listed properties, i.e., \mathbb{Y} and \mathbb{Y}^* are strictly convex and reflexive Banach spaces. This particular case will be referred to as the *reflexive smooth setting*. Two important straightforward consequences need to be remarked in this situation:

- (i) The duality maps $J_{\mathbb{Y}} : \mathbb{Y} \rightarrow \mathbb{Y}^*$ and $J_{\mathbb{Y}^*} : \mathbb{Y}^* \rightarrow \mathbb{Y}^{**}$ are bijective.
- (ii) $J_{\mathbb{Y}^*} = \mathcal{I}_{\mathbb{Y}} \circ J_{\mathbb{Y}}^{-1}$, where $\mathcal{I}_{\mathbb{Y}} : \mathbb{Y} \rightarrow \mathbb{Y}^{**}$ is the canonical injection. Shortly, $J_{\mathbb{Y}^*} = J_{\mathbb{Y}}^{-1}$, by means of canonical identification.

4.3 Subdifferential characterization and minimization

A very important characterization of the duality mapping is given through a *subdifferential*. Let us recall that for a Banach space \mathbb{Y} and function $f : \mathbb{Y} \rightarrow \mathbb{R}$, the subdifferential $\partial f(y)$ of f at a point $y \in \mathbb{Y}$ is defined as the set:

$$\partial f(y) := \left\{ y^* \in \mathbb{Y}^* : f(z) - f(y) \geq \langle y^*, z - y \rangle_{\mathbb{Y}^*, \mathbb{Y}}, \forall z \in \mathbb{Y} \right\}.$$

Proposition 4.4 *Let $f_{\mathbb{Y}} : \mathbb{Y} \rightarrow \mathbb{R}$ be defined by $f_{\mathbb{Y}}(\cdot) := \frac{1}{2} \|\cdot\|_{\mathbb{Y}}^2$. Then, for any $y \in \mathbb{Y}$, the following characterization of the duality mapping holds:*

$$\mathcal{J}_{\mathbb{Y}}(y) = \partial f_{\mathbb{Y}}(y).$$

Proof See, e.g., [7, p. 26]. ■

Remark 4.5 According to Section 4.2.1, the subdifferential of $f_{\mathbb{Y}}(\cdot) = \frac{1}{2} \|\cdot\|_{\mathbb{Y}}^2$ contains exactly one point if \mathbb{Y}^* is strictly convex. In that case, $f_{\mathbb{Y}}$ is Gâteaux differentiable⁴ (see [7, Corollary 2.7]) and for any $y \in \mathbb{Y}$ we have:

$$J_{\mathbb{Y}}(y) = \nabla f_{\mathbb{Y}}(y). \quad \square$$

⁴If \mathbb{Y}^* is uniformly convex, then $f_{\mathbb{Y}}$ is Fréchet differentiable and the duality map $J_{\mathbb{Y}} : \mathbb{Y} \rightarrow \mathbb{Y}^*$ is uniformly continuous on the unit sphere of \mathbb{Y} . Moreover, by the Milman-Pettis Theorem [3], uniform convexity of \mathbb{Y}^* implies reflexivity of \mathbb{Y}^* (hence, reflexivity of \mathbb{Y}).

Example 4.6 (The L^p case) We recall here an explicit formula for the duality map in the Banach space L^p . For $p \in (1, +\infty)$ the space L^p is reflexive and strictly convex (as well as the dual space $L^{p'}$, where $p' = \frac{p}{p-1}$). For $v \in L^p$, the duality map is defined by the action:

$$\langle J_{L^p}(v), w \rangle_{(L^p)^*, L^p} := \|v\|_{L^p}^{2-p} \int |v|^{p-1} \operatorname{sgn}(v) w, \quad \forall w \in L^p,$$

which can be verified by computing the Gâteaux derivative of $v \mapsto (\int |v|^p)^{1/p}$. Observe that $\|v\|_{L^p}^{2-p} |v|^{p-1} \operatorname{sgn}(v) \in L^{p'}$. Moreover, in the case $p = 1$, this formula also works and defines an element in $\mathcal{J}_{L^1}(v)$. Note that L^1 is not a special Banach space as discussed above. \square

The subdifferential is essential in characterizing minimizers. Therefore, as may be expected, the duality mapping naturally appears in the characterization of best approximations.

Theorem 4.A *Let \mathbb{Y} be a Banach space, $\mathbb{M} \subset \mathbb{Y}$ a closed subspace and $y \in \mathbb{Y}$. The following statements are equivalent:*

- (i) $y_0 = \arg \min_{z_0 \in \mathbb{M}} \|y - z_0\|_{\mathbb{Y}}$.
- (ii) *There exists a functional $y^* \in \mathcal{J}_{\mathbb{Y}}(y - y_0)$ which annihilates \mathbb{M} , i.e., $\langle y^*, z_0 \rangle_{\mathbb{Y}^*, \mathbb{Y}} = 0$, for all $z_0 \in \mathbb{M}$.* \square

Proof See, e.g., [18], for the proof in case of $y \in \mathbb{Y} \setminus \mathbb{M}$. The case of $y \in \mathbb{M}$ is a trivial consequence. \blacksquare

An important consequence of Theorem 4.A and the use of the duality mapping, is that we can improve the constant 2 in the standard a priori estimate ($\|y_0\|_{\mathbb{Y}} \leq 2\|y\|_{\mathbb{Y}}$) for the minimizer given in Lemma 3.2. This sharpened estimate seems to be a novel result, which is of independent interest.

Proposition 4.7 *Under the conditions of Theorem 4.A, assume that Theorem 4.A(i) holds true. Then the minimizer $y_0 \in \mathbb{M}$ satisfies the a priori estimate*

$$\|y_0\|_{\mathbb{Y}} \leq (1 + \Lambda_{\mathbb{Y}}) \|y\|_{\mathbb{Y}}, \quad (19)$$

where the geometrical constant $\Lambda_{\mathbb{Y}} \in [0, 1]$ is defined by

$$\Lambda_{\mathbb{Y}} := \sup_{\substack{(z_0, z) \in \mathcal{S}_{\mathbb{Y}} \\ z_0^* \in \mathcal{J}_{\mathbb{Y}}(z_0)}} \frac{\langle z_0^*, z \rangle_{\mathbb{Y}^*, \mathbb{Y}}}{\|z\|_{\mathbb{Y}} \|z_0\|_{\mathbb{Y}}}, \quad (20)$$

and the above supremum is taken over the set $\mathcal{S}_{\mathbb{Y}}$ consisting of all pairs (z_0, z) which are “orthogonal” in the following sense:

$$\mathcal{S}_{\mathbb{Y}} := \left\{ (z_0, z) \in \mathbb{Y} \times \mathbb{Y} : \exists z^* \in \mathcal{J}_{\mathbb{Y}}(z) \text{ satisfying } \langle z^*, z_0 \rangle_{\mathbb{Y}^*, \mathbb{Y}} = 0 \right\}. \quad \square$$

Proof If $y_0 = 0$ or $y_0 = y$, then the result is obvious. Hence, let us assume that $\|y_0\|_{\mathbb{Y}} > 0$ and $\|y - y_0\|_{\mathbb{Y}} > 0$. First of all, we estimate the error using the minimizing property of $y_0 \in \mathbb{M}$:

$$\|y - y_0\|_{\mathbb{Y}} \leq \|y - 0\|_{\mathbb{Y}} = \|y\|_{\mathbb{Y}}. \quad (21)$$

Next, by Theorem 4.A, there exists a $z^* \in \mathcal{J}_{\mathbb{Y}}(y - y_0)$ which annihilates \mathbb{M} . Therefore, $(y_0, y - y_0) \in \mathcal{S}_{\mathbb{Y}}$, and we thus obtain for any $z_0^* \in \mathcal{J}_{\mathbb{Y}}(y_0)$:

$$\begin{aligned} \|y_0\|_{\mathbb{Y}} &= \frac{\langle z_0^*, y_0 \rangle_{\mathbb{Y}^*, \mathbb{Y}}}{\|y_0\|_{\mathbb{Y}}} \\ &= \frac{\langle z_0^*, y \rangle_{\mathbb{Y}^*, \mathbb{Y}}}{\|y_0\|_{\mathbb{Y}}} - \frac{\langle z_0^*, y - y_0 \rangle_{\mathbb{Y}^*, \mathbb{Y}}}{\|y_0\|_{\mathbb{Y}}} \\ &\leq \|y\|_{\mathbb{Y}} - \frac{\langle z_0^*, y - y_0 \rangle_{\mathbb{Y}^*, \mathbb{Y}}}{\|y_0\|_{\mathbb{Y}} \|y - y_0\|_{\mathbb{Y}}} \|y - y_0\|_{\mathbb{Y}} \\ &\leq \|y\|_{\mathbb{Y}} + \Lambda_{\mathbb{Y}} \|y - y_0\|_{\mathbb{Y}}, \end{aligned}$$

Conclude by using the estimate in (21). ■

Remark 4.8 The constant $\Lambda_{\mathbb{Y}}$ is referred to as a “geometric constant” since it measures the degree to which $\langle z_0^*, z \rangle_{\mathbb{Y}^*, \mathbb{Y}}$ fails to be zero, for every $z_0^* \in \mathcal{J}_{\mathbb{Y}}(z_0)$, whenever there is a $z^* \in \mathcal{J}_{\mathbb{Y}}(z)$ such that $\langle z^*, z_0 \rangle_{\mathbb{Y}^*, \mathbb{Y}} = 0$. If \mathbb{Y} is a Hilbert space, then $\Lambda_{\mathbb{Y}} = 0$, since the single-valued duality map $J_{\mathbb{Y}}(\cdot)$ coincides with the self-adjoint Riesz map, and $\langle J_{\mathbb{Y}}(\cdot), \cdot \rangle_{\mathbb{Y}^*, \mathbb{Y}}$ coincides with the inner product in \mathbb{Y} . The maximal value $\Lambda_{\mathbb{Y}} = 1$ holds for example for $\mathbb{Y} = \ell^1(\mathbb{R}^2)$. Indeed taking $z_0 = (1, -1)$ and $z = (\alpha, 1)$, with $\alpha > 0$, then $(2, -2) \in \mathcal{J}_{\mathbb{Y}}(z_0)$ and $(1 + \alpha, 1 + \alpha) \in \mathcal{J}_{\mathbb{Y}}(z)$, so that upon taking $\alpha \rightarrow +\infty$ one obtains $\langle z_0^*, z \rangle_{\mathbb{Y}^*, \mathbb{Y}} / (\|z_0\|_{\mathbb{Y}} \|z\|_{\mathbb{Y}}) \rightarrow 1$. □

Remark 4.9 The result in Proposition 4.7 can also be used to sharpen the a priori bound (14) in Theorem 3.A(ii). □

5 Nonlinear Petrov–Galerkin method and monotone mixed formulation

In this section, we characterize the solution of residual minimization in (10) by means of the duality mapping. The characterization will give rise to a nonlinear Petrov–Galerkin method and corresponding mixed formulation. The practical inexact version of this method is the subject of Section 6.

5.1 Characterization of residual minimization

Theorem 5.A *Let \mathbb{U} and \mathbb{V} be two Banach spaces and let $B : \mathbb{U} \rightarrow \mathbb{V}^*$ be a linear, continuous and bounded-below operator. Given $f \in \mathbb{V}^*$ and a finite-dimensional subspace $\mathbb{U}_n \subset \mathbb{U}$, an element $u_n \in \mathbb{U}_n$ is a solution of the residual minimization problem (10), if and only if there is an $r^{**} \in \mathcal{J}_{\mathbb{V}^*}(f - Bu_n) \subset \mathbb{V}^{**}$ satisfying:*

$$\langle r^{**}, Bw_n \rangle_{\mathbb{V}^{**}, \mathbb{V}^*} = 0, \quad \forall w_n \in \mathbb{U}_n. \quad (22)$$

Proof Apply Theorem 4.A to the minimization problem (10), using $\mathbb{Y} = \mathbb{V}^*$ and $\mathbb{M} = B\mathbb{U}_n$. ■

Defining the discrete space $\mathbb{V}_n^* := B\mathbb{U}_n$, we see that (22) can be interpreted as the nonlinear Petrov–Galerkin method:

$$\begin{aligned} &\text{Find } u_n \in \mathbb{U}_n \text{ such that for some } r^{**} \in \mathcal{J}_{\mathbb{V}^*}(f - Bu_n), \\ &\langle r^{**}, \nu_n \rangle_{\mathbb{V}^{**}, \mathbb{V}^*} = 0, \quad \forall \nu_n \in \mathbb{V}_n^*. \end{aligned} \quad (23)$$

Observe that $r^{**} \in \mathcal{J}_{\mathbb{V}^*}(f - Bu_n)$ must fulfill the set of *nonlinear* equations:

$$\|r^{**}\|_{\mathbb{V}^{**}}^2 = \langle r^{**}, f - Bu_n \rangle_{\mathbb{V}^{**}, \mathbb{V}^*} = \|f - Bu_n\|_{\mathbb{V}^*}^2.$$

Because of specific properties of $\mathcal{J}_{\mathbb{V}^*}$ depending on the geometry of \mathbb{V} , the above characterizations can be reduced to other forms. For example, if \mathbb{V} is reflexive, then $f - Bu_n \in \mathcal{J}_{\mathbb{V}}(r)$ for some $r \in \mathbb{V}$ such that $\langle Bw_n, r \rangle_{\mathbb{V}^*, \mathbb{V}} = 0$, for all $w_n \in \mathbb{U}_n$. To have more useful characterizations, we have to restrict to the reflexive smooth setting.

5.2 The reflexive smooth setting

We apply the previous result to the *reflexive smooth setting* (cf. Section 4.2.4), which refers to the particular situation for which \mathbb{V} and \mathbb{V}^* are strictly convex and reflexive. Many equivalent formulations follow because of this additional structure. Recall that in this setting the duality mapping in \mathbb{V} is a single-valued and bijective map, denoted by $J_{\mathbb{V}} : \mathbb{V} \rightarrow \mathbb{V}^*$.

Theorem 5.B *Under the same conditions of Theorem 5.A, assume additionally that \mathbb{V} and \mathbb{V}^* are strictly convex and reflexive. The following statements are equivalent:*

- (i) $u_n \in \mathbb{U}_n$ is the unique residual minimizer satisfying (10).

(ii) $u_n \in \mathbb{U}_n$ is the solution of the nonlinear Petrov–Galerkin formulation:

$$\langle Bw_n, J_{\mathbb{V}}^{-1}(f - Bu_n) \rangle_{\mathbb{V}^*, \mathbb{V}} = 0, \quad \forall w_n \in \mathbb{U}_n.$$

(iii) There is a unique residual representation $r \in \mathbb{V}$ such that $u_n \in \mathbb{U}_n$ together with r satisfy the semi-infinite monotone mixed formulation:

$$\langle J_{\mathbb{V}}(r), v \rangle_{\mathbb{V}^*, \mathbb{V}} + \langle Bu_n, v \rangle_{\mathbb{V}^*, \mathbb{V}} = \langle f, v \rangle_{\mathbb{V}^*, \mathbb{V}}, \quad \forall v \in \mathbb{V}, \quad (24a)$$

$$\langle B^*r, w_n \rangle_{\mathbb{U}^*, \mathbb{U}} = 0, \quad \forall w_n \in \mathbb{U}_n. \quad (24b)$$

(iv) $u_n \in \mathbb{U}_n$ is the Lagrange multiplier of the constrained minimization:

$$\min_{v \in (B\mathbb{U}_n)^\perp} \frac{1}{2} \|v\|_{\mathbb{V}}^2 - \langle f, v \rangle_{\mathbb{V}^*, \mathbb{V}}. \quad (25)$$

Remark 5.1 Note that for all the formulations in Theorem 5.B, the a priori estimate in (14), its sharpened version (cf. Remark 4.9) and error estimates in (15) are valid, on account of the stated equivalences. \square

Proof

(i) \Leftrightarrow (ii) : By Theorem 5.A, $u_n \in \mathbb{U}_n$ is the unique minimizer of (10) if and only if $J_{\mathbb{V}^*}(f - Bu_n) \in \mathbb{V}^{**}$ annihilates the discrete space $B\mathbb{U}_n \subset \mathbb{V}^*$. In other words, because of the identification $J_{\mathbb{V}^*} = J_{\mathbb{V}}^{-1}$ (see Section 4.2.4),

$$\langle Bw_n, J_{\mathbb{V}}^{-1}(f - Bu_n) \rangle_{\mathbb{V}^*, \mathbb{V}} = \langle J_{\mathbb{V}^*}(f - Bu_n), Bw_n \rangle_{\mathbb{V}^{**}, \mathbb{V}^*} = 0,$$

for all $w_n \in \mathbb{U}_n$.

(ii) \Leftrightarrow (iii) : Note that $r = J_{\mathbb{V}}^{-1}(f - Bu_n)$.

(iii) \Rightarrow (iv) : The Lagrangian $\mathcal{L} : \mathbb{V} \times \mathbb{U}_n \rightarrow \mathbb{R}$ associated with the constrained minimization (25) is:

$$\mathcal{L}(v, w_n) := \frac{1}{2} \|v\|_{\mathbb{V}}^2 - \langle f, v \rangle_{\mathbb{V}^*, \mathbb{V}} + \langle B^*v, w_n \rangle_{\mathbb{U}^*, \mathbb{U}}.$$

Let (r, u_n) denote the solution to the mixed formulation (24). Firstly, since $r \in (B\mathbb{U}_n)^\perp$, it is straightforward to see that $\mathcal{L}(r, w_n) = \mathcal{L}(r, u_n)$. Secondly,

$$\begin{aligned} 0 &= \langle J_{\mathbb{V}}(r), v - r \rangle_{\mathbb{V}^*, \mathbb{V}} + \langle Bu_n, v - r \rangle_{\mathbb{V}^*, \mathbb{V}} - \langle f, v - r \rangle_{\mathbb{V}^*, \mathbb{V}} && \text{(by (24a))} \\ &\leq \frac{1}{2} \|v\|_{\mathbb{V}}^2 - \frac{1}{2} \|r\|_{\mathbb{V}}^2 + \langle Bu_n, v - r \rangle_{\mathbb{V}^*, \mathbb{V}} - \langle f, v - r \rangle_{\mathbb{V}^*, \mathbb{V}} && \text{(by Prop. (4.4))} \\ &= \mathcal{L}(v, u_n) - \mathcal{L}(r, u_n). \end{aligned}$$

Therefore (r, u_n) is a saddle-point of the Lagrangian, i.e.,

$$\mathcal{L}(r, w_n) \leq \mathcal{L}(r, u_n) \leq \mathcal{L}(v, u_n) \quad \forall (v, w_n) \in \mathbb{V} \times \mathbb{U}_n, \quad (26)$$

which is equivalent to (25).

(iv) \Rightarrow (iii) : Let (r, u_n) be a solution of (25), i.e., (26) holds. The first inequality in (26) implies

$$\langle Bw_n, r \rangle_{\mathbb{V}^*, \mathbb{V}} \leq \langle Bu_n, r \rangle_{\mathbb{V}^*, \mathbb{V}} \quad \forall w_n \in \mathbb{U}_n$$

which implies (24b) by a vector-space argument. Next, considering the second inequality in (26) with v equal to $r + \lambda v$, and $\lambda > 0$, it follows that

$$\begin{aligned} 0 &\leq \lambda^{-1} \left(\mathcal{L}(r + \lambda v, u_n) - \mathcal{L}(r, u_n) \right) \\ &= \lambda^{-1} \left(\frac{1}{2} \|r + \lambda v\|_{\mathbb{V}}^2 - \frac{1}{2} \|r\|_{\mathbb{V}}^2 \right) - \langle f, v \rangle_{\mathbb{V}^*, \mathbb{V}} + \langle Bu_n, v \rangle_{\mathbb{V}^*, \mathbb{V}} \\ &\leq \langle J_{\mathbb{V}}(r + \lambda v), v \rangle_{\mathbb{V}^*, \mathbb{V}} + \langle Bu_n, v \rangle_{\mathbb{V}^*, \mathbb{V}} - \langle f, v \rangle_{\mathbb{V}^*, \mathbb{V}} \quad (\text{by Prop. 4.4}) \end{aligned}$$

Therefore, upon $\lambda \rightarrow 0^+$, invoking hemi-continuity of $J_{\mathbb{V}}$ (see (17)) and repeating the above with $-v$ instead of v , one recovers (24a). \blacksquare

Remark 5.2 If one assumes B is bijective and, instead of $\|\cdot\|_{\mathbb{V}}$, one uses the norm $\|\cdot\|_{\mathbb{V}_{\text{opt}}}$ on \mathbb{V} (recall from Remark 3.5), then one can show that (24a) holds with $\langle J_{\mathbb{V}}(r), v \rangle_{\mathbb{V}^*, \mathbb{V}}$ replaced by $\langle B^*v, J_{\mathbb{U}}^{-1}(B^*r) \rangle_{\mathbb{U}^*, \mathbb{U}}$. \square

6 The inexact monotone mixed method

We now consider a *tractable* approximation to our optimal Petrov–Galerkin discretization in the reflexive smooth setting (cf. Section 4.2.4). As usual, we consider two Banach spaces \mathbb{U} and \mathbb{V} , together with a linear, continuous and bounded-below operator $B : \mathbb{U} \rightarrow \mathbb{V}^*$. The reflexive smooth setting guarantees that the semi-infinite mixed formulation (24) introduced in Section 5.2 is well posed. This formulation will be the starting point for our inexact method.

In addition to $\mathbb{U}_n \subset \mathbb{U}$, let $\mathbb{V}_m \subset \mathbb{V}$ be a finite-dimensional subspace. We shall then consider the following inexact approximation to (24), which is a fully-discrete mixed method:

Find $(r_m, u_n) \in \mathbb{V}_m \times \mathbb{U}_n$ such that

$$\langle J_{\mathbb{V}}(r_m), v_m \rangle_{\mathbb{V}^*, \mathbb{V}} + \langle Bu_n, v_m \rangle_{\mathbb{V}^*, \mathbb{V}} = \langle f, v_m \rangle_{\mathbb{V}^*, \mathbb{V}} \quad \forall v_m \in \mathbb{V}_m, \quad (27a)$$

$$\langle B^*r_m, w_n \rangle_{\mathbb{U}^*, \mathbb{U}} = 0 \quad \forall w_n \in \mathbb{U}_n. \quad (27b)$$

Because the nonlinearity caused by $J_{\mathbb{V}}$ is monotone, we refer to the above as a monotone mixed method.

6.1 Equivalent discrete settings

Analogous to the semi-infinite mixed formulation, which is equivalent to a residual minimization (see Theorem 5.B), the inexact monotone mixed method is also related to a particular minimization of the residual, taken now with respect to the *discrete* dual norm

$$\|\cdot\|_{(\mathbb{V}_m)^*} = \sup_{v_m \in \mathbb{V}_m} \frac{\langle \cdot, v_m \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m}}{\|v_m\|_{\mathbb{V}}},$$

The next theorem summarises this equivalence and, additionally, shows the equivalence with a *discrete* constrained minimization.

Theorem 6.A *Let \mathbb{U} and \mathbb{V} be two Banach spaces and let $B : \mathbb{U} \rightarrow \mathbb{V}^*$ be a linear, continuous and bounded-below operator. Assume that \mathbb{V} and \mathbb{V}^* are reflexive and strictly convex. Given $f \in \mathbb{V}^*$ and finite-dimensional subspaces $\mathbb{U}_n \subset \mathbb{U}$ and $\mathbb{V}_m \subset \mathbb{V}$, the following statements are equivalent:*

- (i) $(u_n, r_m) \in \mathbb{U}_n \times \mathbb{V}_m$ is a solution of the discrete mixed problem (27).
- (ii) $u_n \in \mathbb{U}_n$ is a minimizer of the discrete residual minimization problem:

$$\min_{w_n \in \mathbb{U}_n} \|I_m^*(f - Bw_n)\|_{(\mathbb{V}_m)^*}, \quad (28)$$

and $r_m = J_{\mathbb{V}_m}^{-1} \circ I_m^*(f - Bu_n)$, with $I_m : \mathbb{V}_m \rightarrow \mathbb{V}$ the natural injection.

- (iii) $u_n \in \mathbb{U}_n$ is the Lagrange multiplier of the discrete constrained minimization problem:

$$\min_{v_m \in \mathbb{V}_m \cap (B\mathbb{U}_n)^\perp} \left\{ \frac{1}{2} \|v_m\|_{\mathbb{V}}^2 - \langle f, v_m \rangle_{\mathbb{V}^*, \mathbb{V}} \right\}, \quad (29)$$

while $r_m \in \mathbb{V}_m$ is the minimizer of it. □

Proof We demonstrate each equivalence with respect to (i).

(i) \Leftrightarrow (ii) : First we note the following direct equivalences:

$$\begin{aligned} r_m &= J_{\mathbb{V}_m}^{-1} \circ I_m^*(f - Bu_n) \\ \Leftrightarrow J_{\mathbb{V}_m}(r_m) &= I_m^*(f - Bu_n) \\ \Leftrightarrow I_m^* J_{\mathbb{V}}(I_m r_m) &= I_m^*(f - Bu_n), \end{aligned} \quad (\text{by Lemma 4.3})$$

while the last statement is equivalent to (27a).

Next, if $(u_n, r_m) \in \mathbb{U}_n \times \mathbb{V}_m$ is a solution of (27), then for any $w_n \in \mathbb{U}_n$ we have:

$$\begin{aligned}
\|I_m^*(f - Bu_n)\|_{(\mathbb{V}_m)^*} &= \sup_{v_m \in \mathbb{V}_m} \frac{\langle I_m^*(f - Bu_n), v_m \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m}}{\|v_m\|_{\mathbb{V}}} \\
&= \sup_{v_m \in \mathbb{V}_m} \frac{\langle f - Bu_n, I_m v_m \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|v_m\|_{\mathbb{V}}} \\
&= \sup_{v_m \in \mathbb{V}_m} \frac{\langle J_{\mathbb{V}}(r_m), v_m \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|v_m\|_{\mathbb{V}}} && \text{(by (27a))} \\
&= \frac{\langle J_{\mathbb{V}}(r_m), r_m \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|r_m\|_{\mathbb{V}}} && \text{(by (16))} \\
&= \frac{\langle f - Bu_n, r_m \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|r_m\|_{\mathbb{V}}} && \text{(by (27a))} \\
&= \frac{\langle f - Bw_n, r_m \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|r_m\|_{\mathbb{V}}} && \text{(by (27b))} \\
&= \frac{\langle I_m^*(f - Bw_n), r_m \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m}}{\|r_m\|_{\mathbb{V}}} \\
&\leq \|I_m^*(f - Bw_n)\|_{(\mathbb{V}_m)^*}
\end{aligned}$$

Thus, u_n is a minimizer of (28).

Finally, if $u_n \in \mathbb{U}_n$ is a minimizer of (28) and $r_m = J_{\mathbb{V}_m}^{-1} \circ I_m^*(f - Bu_n) = J_{(\mathbb{V}_m)^*} \circ I_m^*(f - Bu_n)$, then by Theorem 4.A, with $\mathbb{Y} = (\mathbb{V}_m)^*$ and $\mathbb{M} = I_m^* B \mathbb{U}_n \subset (\mathbb{V}_m)^*$, r_m satisfies

$$0 = \langle I_m^* B w_n, r_m \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m} = \langle B w_n, I_m r_m \rangle_{\mathbb{V}^*, \mathbb{V}} = \langle B^* r_m, w_n \rangle_{\mathbb{U}^*, \mathbb{U}}, \quad \forall w_n \in \mathbb{U}_n,$$

which verifies (27b).

(i) \Leftrightarrow (iii) : The proof of this equivalence follows exactly the same reasoning as in the semi-infinite setting; see the proof of Theorem 5.B, part (iii) \Leftrightarrow (iv). ■

6.2 Well-posedness of the inexact method

We now study the existence and uniqueness of solutions to the inexact monotone mixed method (27). A critical ingredient for the uniqueness analysis is the existence of an operator $\Pi : \mathbb{V} \rightarrow \mathbb{V}_m$ such that:

$$\|\Pi v\|_{\mathbb{V}} \leq C_{\Pi} \|v\|_{\mathbb{V}}, \quad \forall v \in \mathbb{V} \text{ and some constant } C_{\Pi} > 0; \quad (30a)$$

$$\|(I - \Pi)v\|_{\mathbb{V}} \leq D_{\Pi} \|v\|_{\mathbb{V}}, \quad \forall v \in \mathbb{V} \text{ and some constant } D_{\Pi} > 0; \quad (30b)$$

$$\langle B w_n, v - \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}} = 0, \quad \forall w_n \in \mathbb{U}_n, \forall v \in \mathbb{V}, \quad (30c)$$

where $I : \mathbb{V} \rightarrow \mathbb{V}$ is the identity map in \mathbb{V} . Such an operator is referred to as a *Fortin operator* after Fortin's trick in mixed finite element methods [1, Section 5.4].

For the existence of Π , note that the last identity (30c) requires that $\dim \mathbb{V}_m \geq \dim \operatorname{Im}(B) = \dim \mathbb{U}_n$ (for a bounded-below operator B).

Theorem 6.B *Under the same conditions of Theorem 6.A, assume additionally that an operator $\Pi : \mathbb{V} \rightarrow \mathbb{V}_m$ satisfying (30) exists. Then, the inexact monotone mixed method (27) has a unique solution $(r_m, u_n) \in \mathbb{V}_m \times \mathbb{U}_n$. \square*

Note that by the equivalences in Theorem 6.A, Theorem 6.B also implies that the discrete residual minimization and discrete constrained minimization problems are well-posed.

Proof To proof existence, we consider the equivalent discrete constrained minimization problem (29). The existence of a minimizer $r_m \in \mathbb{V}_m \cap (B\mathbb{U}_n)^\perp$ is guaranteed since the functional $v_m \mapsto \frac{1}{2}\|v_m\|_{\mathbb{V}}^2 - \langle f, v_m \rangle_{\mathbb{V}^*, \mathbb{V}}$ is convex and continuous, and $\mathbb{V}_m \cap (B\mathbb{U}_n)^\perp$ is a closed subspace.

Next, we claim that there exist a $u_n \in \mathbb{U}_n$ such that

$$\langle Bu_n, v_m \rangle_{\mathbb{V}^*, \mathbb{V}} = \langle f - J_{\mathbb{V}}(r_m), v_m \rangle_{\mathbb{V}^*, \mathbb{V}}, \quad \forall v_m \in \mathbb{V}_m.$$

To see this, consider the restricted operator $B_n : \mathbb{U}_n \rightarrow \mathbb{V}^*$, such that $B_n w_n = Bw_n$ for all $w_n \in \mathbb{U}_n$, and recall the natural injection $I_m : \mathbb{V}_m \rightarrow \mathbb{V}$. Then, the above translates into

$$I_m^* B_n u_n = I_m^* (f - J_{\mathbb{V}}(r_m)) \quad \text{in } (\mathbb{V}_m)^*.$$

Thus, to proof existence, we show that $I_m^* (f - J_{\mathbb{V}}(r_m))$ is in the (closed) range of the finite-dimensional operator $I_m^* B_n : \mathbb{U}_n \rightarrow (\mathbb{V}_m)^*$. Since r_m is the minimizer of (29), we have

$$\begin{aligned} 0 &= \langle J_{\mathbb{V}}(r_m) - f, I_m v_m \rangle_{\mathbb{V}^*, \mathbb{V}} = \langle I_m^* (J_{\mathbb{V}}(r_m) - f), v_m \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m}, \\ &\forall v_m \in \mathbb{V}_m \cap (B\mathbb{U}_n)^\perp = \operatorname{Ker}(B_n^* I_m). \end{aligned}$$

Hence, $I_m^* (f - J_{\mathbb{V}}(r_m)) \in (\operatorname{Ker}(B_n^* I_m))^\perp = \operatorname{Im}(I_m^* B_n)$.

To prove uniqueness assume that (u_n, r_m) and $(\tilde{u}_n, \tilde{r}_m)$ are two solutions of problem (27). Then, by subtraction, it is immediate to see that:

$$\langle J_{\mathbb{V}}(r_m) - J_{\mathbb{V}}(\tilde{r}_m), r_m - \tilde{r}_m \rangle_{\mathbb{V}^*, \mathbb{V}} = 0,$$

which implies that $\tilde{r}_m = r_m$ by strict monotonicity of $J_{\mathbb{V}}$ (see (18)). Going back to (27a) we now obtain $\langle B(u_n - \tilde{u}_n), v_m \rangle_{\mathbb{V}^*, \mathbb{V}} = 0$, for all $v_m \in \mathbb{V}_m$. Therefore, by the Fortin-operator property (30c),

$$\langle B(u_n - \tilde{u}_n), v \rangle_{\mathbb{V}^*, \mathbb{V}} = \langle B(u_n - \tilde{u}_n), \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}} = 0, \quad \forall v \in \mathbb{V}.$$

Thus, $B(u_n - \tilde{u}_n) = 0$ which implies $u_n - \tilde{u}_n = 0$ since B is bounded below. \blacksquare

6.3 Error analysis of the inexact method

We next present an error analysis for the inexact monotone mixed method. Since the method is fundamentally related to (discrete) residual minimization, the most straightforward error estimate is of *a posteriori* type. Immediately after, an *a priori* error estimate follows naturally from the *a posteriori* estimate (compare with the error estimates for the *exact* residual-minimization method in (15)). The constant in the resulting *a priori* estimate can however be improved by resorting to an alternative analysis technique, which we present in Section 6.4.

In the following results, recall that $M_B > 0$ and $\gamma_B > 0$ denote the continuity and bounded-below constants of the operator $B : \mathbb{U} \rightarrow \mathbb{V}^*$, and that C_Π and D_Π are the constants in (30a)–(30b) related to the Fortin operator $\Pi : \mathbb{V} \rightarrow \mathbb{V}_m$.

Theorem 6.C *Under the same conditions of Theorem 6.B, let $(u_n, r_m) \in \mathbb{U}_n \times \mathbb{V}_m$ be the unique solution of the discrete mixed problem (27). If $f \in \text{Im}(B)$ and $u \in \mathbb{U}$ is the solution of the continuous problem $Bu = f$, then u_n satisfies the following *a posteriori* error estimate:*

$$\|u - u_n\|_{\mathbb{U}} \leq \frac{1}{\gamma_B} \text{osc}(f) + \frac{C_\Pi}{\gamma_B} \|r_m\|_{\mathbb{V}}, \quad (31)$$

where the data-oscillation term $\text{osc}(f)$ satisfies

$$\text{osc}(f) := \sup_{v \in \mathbb{V}} \frac{\langle f, v - \Pi v \rangle}{\|v\|_{\mathbb{V}}} \leq M_B D_\Pi \inf_{w_n \in \mathbb{U}_n} \|u - w_n\|_{\mathbb{U}}, \quad (32)$$

and r_m satisfies

$$\|r_m\|_{\mathbb{V}} = \|I_m^*(f - Bu_n)\|_{(\mathbb{V}_m)^*} \leq M_B \inf_{w_n \in \mathbb{U}_n} \|u - w_n\|_{\mathbb{U}}. \quad (33)$$

Remark 6.1 The *a posteriori* error estimate in Theorem 6.C extends the result by Carstensen, Demkowicz and Gopalakrishnan [5] for the Hilbert-space version of the method. \square

Proof Using that the operator B is bounded from below, and that $Bu = f$, we get:

$$\|u - u_n\|_{\mathbb{U}} \leq \frac{1}{\gamma_B} \|Bu - Bu_n\|_{\mathbb{V}^*} = \frac{1}{\gamma_B} \sup_{v \in \mathbb{V}} \frac{\langle f - Bu_n, v - \Pi v + \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|v\|_{\mathbb{V}}}.$$

Next, by definition of the Π operator (eq. (30)), $Bu_n \in \mathbb{V}^*$ annihilates $v - \Pi v$, for all $v \in \mathbb{V}$. Hence, splitting the supremum we obtain:

$$\begin{aligned} \|u - u_n\|_{\mathbb{U}} &\leq \frac{1}{\gamma_B} \sup_{v \in \mathbb{V}} \frac{\langle f, v - \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|v\|_{\mathbb{V}}} + \frac{1}{\gamma_B} \sup_{v \in \mathbb{V}} \frac{\langle f - Bu_n, \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|v\|_{\mathbb{V}}} \\ &\leq \frac{1}{\gamma_B} \text{osc}(f) + \frac{C_\Pi}{\gamma_B} \sup_{v \in \mathbb{V}} \frac{\langle f - Bu_n, \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|\Pi v\|_{\mathbb{V}}}, \end{aligned}$$

where we used boundedness of Π . To conclude, it must be observed that:

$$\sup_{v \in \mathbb{V}} \frac{\langle f - Bu_n, \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|\Pi v\|_{\mathbb{V}}} = \sup_{v \in \mathbb{V}} \frac{\langle J_{\mathbb{V}}(r_m), \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|\Pi v\|_{\mathbb{V}}} \leq \|J_{\mathbb{V}}(r_m)\|_{\mathbb{V}^*} = \|r_m\|_{\mathbb{V}}.$$

Next, observe that for all $w_n \in \mathbb{U}_n$ we have

$$\text{osc}(f) = \sup_{v \in \mathbb{V}} \frac{\langle f, v - \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|v\|_{\mathbb{V}}} = \sup_{v \in \mathbb{V}} \frac{\langle f - Bw_h, v - \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|v\|_{\mathbb{V}}} \leq M_B D_{\Pi} \|u - w_h\|_{\mathbb{U}}.$$

Finally, by the proof of Theorem 6.A, part (i) \Leftrightarrow (ii),

$$\|I_m^*(f - Bu_n)\|_{(\mathbb{V}_m)^*} = \frac{\langle J_{\mathbb{V}}(r_m), r_m \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|r_m\|_{\mathbb{V}}} = \|r_m\|_{\mathbb{V}} \leq \|I_m^*(f - Bw_n)\|_{(\mathbb{V}_m)^*},$$

and

$$\|I_m^*(f - Bw_n)\|_{(\mathbb{V}_m)^*} \leq \|Bu - Bw_n\|_{\mathbb{V}^*} \leq M_B \|u - w_n\|_{\mathbb{U}}. \quad \blacksquare$$

A straightforward a priori error estimate follows naturally from the results in Theorem 6.C.

Corollary 6.2 *Under the same assumptions of Theorem 6.C, we have the following a priori error estimate:*

$$\|u - u_n\|_{\mathbb{U}} \leq \frac{1}{\gamma_B} \text{osc}(f) + \frac{C_{\Pi} M_B}{\gamma_B} \inf_{w_n \in \mathbb{U}_n} \|u - w_n\|_{\mathbb{U}} \quad (34a)$$

$$\leq \frac{(D_{\Pi} + C_{\Pi}) M_B}{\gamma_B} \inf_{w_n \in \mathbb{U}_n} \|u - w_n\|_{\mathbb{U}} \quad (34b)$$

Remark 6.3 In the context of finite-element approximations, the data-oscillation term in (34a) can generally be expected to be of higher order than indicated by the upper bound in (34b); see discussion in [5].

Note that if $\mathbb{V}_m = \mathbb{V}$, then $\text{osc}(f) = 0$, $D_{\Pi} = 0$ and $C_{\Pi} = 1$ (choose $\Pi = I$), so that the estimates in (31) and (34) reduce to those in the semi-infinite case (15). \square

6.4 Direct a priori error analysis of the inexact method

A direct a priori error analysis is possible for the inexact method, without going through an a posteriori error estimate. The benefit of the direct analysis is that the resulting estimate is sharper than the worst-case upper bound given in (34b).

The main idea of the direct analysis is based on the sequence of inequalities (formalized below):

$$\|u - u_n\|_{\mathbb{U}} \leq \|I - P_n\| \|u - w_n\|_{\mathbb{U}} \leq C \|P_n\| \|u - w_n\|_{\mathbb{U}} \quad \forall w_n \in \mathbb{U}_n, \quad (35)$$

where I is the identity, P_n is the projector defined below in Definition 6.6 and the norm $\|\cdot\|$ is defined in the following.

Definition 6.4 Let \mathbb{X}, \mathbb{Y} be two normed vector spaces. A nonlinear map $Q : \mathbb{X} \rightarrow \mathbb{Y}$ is said to be *bounded* if there is a constant $C > 0$ such that $\|Q(x)\|_{\mathbb{Y}} \leq C\|x\|_{\mathbb{X}}$, for all $x \in \mathbb{X}$. In that case, the norm of Q is defined by

$$\|Q\| := \sup_{x \in \mathbb{X}} \frac{\|Q(x)\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}}. \quad \square$$

For the second inequality in (35) we extend a recent result by Stern [20, Theorem 3] on Petrov–Galerkin projectors. That result depends on a geometric constant $C_{\text{BM}}(\mathbb{X}) \in [1, 2]$, referred to as the Banach–Mazur constant, which also quantifies how much a Banach space \mathbb{X} fails to be Hilbert (compare with Remark 4.8 on the different geometric constant $\Lambda_{\mathbb{Y}}$). The constant $C_{\text{BM}}(\mathbb{X})$ is defined by

$$C_{\text{BM}}(\mathbb{X}) := \sup \left\{ (d_{\text{BM}}(\mathbb{M}, \ell_2(\mathbb{R}^2)))^2 : \mathbb{M} \subset \mathbb{X}, \dim \mathbb{M} = 2 \right\},$$

where $d_{\text{BM}}(\cdot, \cdot)$ is the Banach–Mazur distance:

$$d_{\text{BM}}(\mathbb{M}, \ell_2(\mathbb{R}^2)) := \inf \left\{ \|T\| \|T^{-1}\| : T \text{ is a linear isomorphism } \mathbb{M} \rightarrow \ell_2(\mathbb{R}^2) \right\}.$$

It is known that $C_{\text{BM}}(\mathbb{X}) = 1$ if and only if \mathbb{X} is a Hilbert space.

Lemma 6.5 *Let \mathbb{X} be a normed space, $I : \mathbb{X} \rightarrow \mathbb{X}$ the identity and $P : \mathbb{X} \rightarrow \mathbb{X}$ a nonlinear operator such that:*

- (i) P is bounded.
- (ii) $0 \neq P = P \circ P \neq I$.
- (iii) $P(\lambda x) = \lambda P(x)$, $\forall x \in \mathbb{X}$ and $\forall \lambda \in \mathbb{R}$.
- (iv) $P(x - P(y)) = P(x) - P(y)$, for all $x, y \in \mathbb{X}$.

Then the nonlinear operator $I - P$ is bounded and satisfies

$$\|I - P\| \leq C_S \|P\|, \quad \square$$

where C_S is the constant introduced by Stern [20]:

$$C_S = \min \left\{ 1 + \|P\|^{-1}, C_{\text{BM}}(\mathbb{X}) \right\}. \quad (36)$$

Proof This result is Theorem 3 in Stern [20]. Although that Theorem considers linear projectors, one can show that it extends to projectors with the properties in (i)–(iv). ■

To define our projector P_n , consider any $u \in \mathbb{U}$. Next, let $(r_m, u_n) \in \mathbb{V}_m \times \mathbb{U}_n$ be the solution of the inexact monotone mixed method (27) with $f = Bu \in \mathbb{V}^*$, i.e.,

$$\langle J_{\mathbb{V}}(r_m), v_m \rangle_{\mathbb{V}^*, \mathbb{V}} + \langle Bu_n, v_m \rangle_{\mathbb{V}^*, \mathbb{V}} = \langle Bu, v_m \rangle_{\mathbb{V}^*, \mathbb{V}} \quad \forall v_m \in \mathbb{V}_m, \quad (37a)$$

$$\langle B^* r_m, w_n \rangle_{\mathbb{U}^*, \mathbb{U}} = 0 \quad \forall w_n \in \mathbb{U}_n. \quad (37b)$$

Definition 6.6 Under the same conditions of Theorem 6.B, we define the *nonlinear Petrov–Galerkin projector* to be the well-defined map

$$P_n : \mathbb{U} \rightarrow \mathbb{U}_n \quad \text{such that} \quad P_n(u) := u_n,$$

with u_n the second argument of the solution (r_m, u_n) of (37). \square

The next result establishes a fundamental bound on P_n .

Proposition 6.7 Under the conditions of Theorem 6.B, let $P_n : \mathbb{U} \rightarrow \mathbb{U}_n$ denote the nonlinear Petrov–Galerkin projector of Definition 6.6. Then P_n is bounded and satisfies the upper bound

$$\|P_n\| = \sup_{u \in \mathbb{U}} \frac{\|P_n(u)\|_{\mathbb{U}}}{\|u\|_{\mathbb{U}}} \leq \frac{C_{\Pi}}{\gamma_B} (1 + \Lambda_{\mathbb{V}}) M_B \quad (38)$$

where the geometric constant $\Lambda_{\mathbb{V}} \in [0, 1]$ is given by

$$\Lambda_{\mathbb{V}} = \sup_{(z, v) \in \mathcal{S}_{\mathbb{V}}} \frac{\langle J_{\mathbb{V}}(z), v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|z\|_{\mathbb{V}} \|v\|_{\mathbb{V}}},$$

with $\mathcal{S}_{\mathbb{V}} = \{(z, v) \in \mathbb{V} \times \mathbb{V} : \langle J_{\mathbb{V}}(v), z \rangle_{\mathbb{V}^*, \mathbb{V}} = 0\}$. \square

Proof Consider any $u \in \mathbb{U}$ and let $(r_m, u_n) \in \mathbb{V}_m \times \mathbb{U}_n$ denote the solution to (37). Using Theorem 6.A we know that $u_n \in \mathbb{U}_n$ corresponds to the discrete residual minimizer

$$u_n = \arg \min_{w_n \in \mathbb{U}_n} \|I_m^*(Bu - Bw_n)\|_{(\mathbb{V}_m)^*}.$$

Applying Proposition 4.7 at the discrete level we get:

$$\|I_m^* Bu_n\|_{(\mathbb{V}_m)^*} \leq (1 + \Lambda_{(\mathbb{V}_m)^*}) \|I_m^* Bu\|_{(\mathbb{V}_m)^*} \leq (1 + \Lambda_{(\mathbb{V}_m)^*}) M_B \|u\|_{\mathbb{U}}, \quad (39)$$

with

$$\Lambda_{(\mathbb{V}_m)^*} = \sup_{(v_m^*, z_m^*) \in \mathcal{S}_{(\mathbb{V}_m)^*}} \frac{\langle z_m^*, J_{\mathbb{V}_m}^{-1}(v_m^*) \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m}}{\|v_m^*\|_{(\mathbb{V}_m)^*} \|z_m^*\|_{(\mathbb{V}_m)^*}}$$

$$\mathcal{S}_{(\mathbb{V}_m)^*} = \left\{ (v_m^*, z_m^*) \in (\mathbb{V}_m)^* \times (\mathbb{V}_m)^* : \langle v_m^*, J_{\mathbb{V}_m}^{-1}(z_m^*) \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m} = 0 \right\},$$

where we have used the identification $J_{(\mathbb{V}_m)^*} = J_{\mathbb{V}_m}^{-1}$ (see Section 4.2.4). To show that $\Lambda_{(\mathbb{V}_m)^*} \leq \Lambda_{\mathbb{V}}$, observe first that, by surjectivity of the duality map $J_{\mathbb{V}_m}$,

$$\begin{aligned} \mathcal{S}_{(\mathbb{V}_m)^*} &= \left\{ (J_{\mathbb{V}_m}(v_m), J_{\mathbb{V}_m}(z_m)) \in (\mathbb{V}_m)^* \times (\mathbb{V}_m)^* : \langle J_{\mathbb{V}_m}(v_m), z_m \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m} = 0, \right. \\ &\quad \left. (z_m, v_m) \in \mathbb{V}_m \times \mathbb{V}_m \right\} \\ &= \left\{ (J_{\mathbb{V}_m}(v_m), J_{\mathbb{V}_m}(z_m)) \in (\mathbb{V}_m)^* \times (\mathbb{V}_m)^* : (z_m, v_m) \in \mathcal{S}_{\mathbb{V}_m} \right\} \end{aligned}$$

and $\mathcal{S}_{\mathbb{V}_m} = \{(z_m, v_m) \in \mathbb{V}_m \times \mathbb{V}_m : \langle J_{\mathbb{V}_m}(v_m), z_m \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m}\}$. Next, noting that $\|J_{\mathbb{V}_m}(\cdot)\|_{(\mathbb{V}_m)^*} = \|\cdot\|_{\mathbb{V}_m}$ and invoking Lemma 4.3, it is now easy to see that:

$$\Lambda_{(\mathbb{V}_m)^*} = \sup_{(z_m, v_m) \in \mathcal{S}_{\mathbb{V}_m}} \frac{\langle J_{\mathbb{V}_m}(z_m), v_m \rangle_{(\mathbb{V}_m)^*, \mathbb{V}_m}}{\|z_m\|_{\mathbb{V}} \|v_m\|_{\mathbb{V}}} \leq \sup_{(z, v) \in \mathcal{S}_{\mathbb{V}}} \frac{\langle J_{\mathbb{V}}(z), v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|z\|_{\mathbb{V}} \|v\|_{\mathbb{V}}} = \Lambda_{\mathbb{V}}, \quad (40)$$

because the last supremum covers a larger set.

On another hand, we use the bounded-belowness of B and Fortin property to obtain:

$$\|u_n\|_{\mathbb{U}} \leq \frac{1}{\gamma_B} \sup_{v \in \mathbb{V}} \frac{\langle Bu_n, v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|v\|_{\mathbb{V}}} \leq \frac{C_{\Pi}}{\gamma_B} \sup_{v \in \mathbb{V}} \frac{\langle Bu_n, \Pi v \rangle_{\mathbb{V}^*, \mathbb{V}}}{\|\Pi v\|_{\mathbb{V}}} \leq \frac{C_{\Pi}}{\gamma_B} \|I_m^* Bu_n\|_{(\mathbb{V}_m)^*}. \quad (41)$$

The result follows by combining inequalities (39), (40) and (41). \blacksquare

We now obtain properties for $(I - P_n)$ by establishing that P_n satisfies all the required properties to invoke Lemma 6.5.

Lemma 6.8 *Let $P_n : \mathbb{U} \rightarrow \mathbb{U}_n$ denote the nonlinear Petrov–Galerkin projector of Definition 6.6 and let $I : \mathbb{U} \rightarrow \mathbb{U}$ be the identity operator. Then, P_n satisfies the assumptions of Lemma 6.5 and therefore*

$$\|I - P_n\| \leq \min \left\{ \frac{C_{\Pi}}{\gamma_B} (1 + \Lambda_{\mathbb{V}}) M_B C_{\text{BM}}(\mathbb{U}), 1 + \frac{C_{\Pi}}{\gamma_B} (1 + \Lambda_{\mathbb{V}}) M_B \right\}. \quad (42)$$

Moreover,

$$(I - P_n)(u) = (I - P_n)(u - w_n), \quad \text{for any } u \in \mathbb{U} \text{ and } w_n \in \mathbb{U}_n. \quad (43)$$

Proof First, we verify one by one that P_n satisfies the assumptions of Lemma 6.5.

(i) Boundedness follows from Proposition 6.7.

(ii) Take $u \in \mathbb{U}$ and plug $u_n = P_n(u)$ in the right-hand side of equation (37a). Then the unique solution of the mixed system (37) will be $(0, u_n)$. Therefore $P_n(P_n(u)) = P_n(u_n) = u_n$. The fact that $P_n \neq 0$ or I is easy to verify whenever $\mathbb{U}_n \neq \{0\}$ or \mathbb{U} .

- (iii) The result follows by multiplying both equations of the mixed system (37) by $\lambda \in \mathbb{R}$ and using the homogeneity of the duality map (see Proposition 4.2).
- (iv) Let (r_m, u_n) be the solution of the mixed system (37) and for some $\phi \in \mathbb{U}$, let $\phi_n = P_n(\phi) \in \mathbb{U}_n$. By subtracting $\langle B\phi_n, v_m \rangle_{\mathbb{V}^*, \mathbb{V}}$ on both sides of the identity in (37a), we get that $(r_m, u_n - \phi_n)$ is the unique solution of (37) with right-hand side $\langle B(u - \phi_n), v_m \rangle_{\mathbb{V}^*, \mathbb{V}}$. Therefore $P(u - \phi_n) = u_n - \phi_n$.

The upper bound (42) is a direct application of Lemma 6.5 combined with (38). The last statement (43) follows straightforwardly from the fact that P_n satisfies assumption (iv) of Lemma (6.5). \blacksquare

We now have all the ingredients for the *sharpened* a priori error estimate.

Theorem 6.D *Under the conditions of Theorem 6.B, let $(u_n, r_m) \in \mathbb{U}_n \times \mathbb{V}_m$ be the unique solution of the inexact method (27). If $f \in \text{Im}(B)$ and $u \in \mathbb{U}$ is the solution of the continuous problem $Bu = f$, then u_n satisfies the a priori error estimate:*

$$\|u - u_n\|_{\mathbb{V}} \leq C \inf_{w_n \in \mathbb{U}_n} \|u - w_n\|_{\mathbb{U}}$$

with

$$C = \min \left\{ \frac{C_{\Pi}}{\gamma_B} (1 + \Lambda_{\mathbb{V}}) M_B C_{\text{BM}}(\mathbb{U}), 1 + \frac{C_{\Pi}}{\gamma_B} (1 + \Lambda_{\mathbb{V}}) M_B \right\}. \quad \square$$

Proof Let $w_n \in \mathbb{U}_n$. By Lemma 6.8 (see eq. (43)), we have

$$\|u - u_n\|_{\mathbb{U}} = \|(I - P_n)u\|_{\mathbb{U}} = \|(I - P_n)(u - w_n)\|_{\mathbb{U}} \leq \|I - P_n\| \|u - w_n\|_{\mathbb{U}}.$$

The result follows by using the upper bound of $\|I - P_n\|$ (see eq. (42)) and taking the infimum of all $w_n \in \mathbb{U}_n$. \blacksquare

Remark 6.9 If \mathbb{U} and \mathbb{V} are Hilbert spaces, then $C_{\text{BM}} = 1$, $\Lambda_{\mathbb{V}} = 0$ and $C = C_{\Pi} M_B / \gamma_B$ in the above a priori error estimate. This recovers the result in the Hilbert-space setting, see, e.g., [11]. \square

Acknowledgements

KGvdZ and IM thank Leszek Demkowicz, Michael Holst and Sarah Pollock for initial discussions on the topic.

- [1] D. BOFFI, F. BREZZI, AND M. FORTIN, *Mixed Finite Element Methods and Applications*, vol. 44 of Springer Series in Computational Mathematics, Springer, Berlin, 2013.
- [2] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, vol. 15 of Texts in Applied Mathematics, Springer, Berlin, 3rd ed., 2008.
- [3] H. BREZIS, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer, New York, 2011.
- [4] D. BROERSEN AND R. STEVENSON, *A robust Petrov–Galerkin discretisation of convection–diffusion equations*, Comput. Math. Appl., 68 (2014), pp. 1605–1618.
- [5] C. CARSTENSEN, L. DEMKOWICZ, AND J. GOPALAKRISHNAN, *A posteriori error control for DPG methods*, SIAM J. Numer. Anal., 52 (2014), pp. 1335–1353.
- [6] C. CHIDUME, *Geometric Properties of Banach Spaces and Nonlinear Iterations*, vol. 1965 of Lecture Notes in Mathematics, Springer, London, 2009.
- [7] I. CIORANESCU, *Geometry of Banach Spaces, Duality Mappings and Nonlinear Problems*, vol. 62 of Mathematics and Its Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.
- [8] A. COHEN, W. DAHMEN, AND G. WELPER, *Adaptivity and variational stabilization for convection-diffusion equations*, M2AN Math. Model. Numer. Anal., 46 (2012), pp. 1247–1273.
- [9] K. DEIMLING, *Nonlinear Functional Analysis*, Springer, Berlin, 1985.
- [10] L. DEMKOWICZ AND J. GOPALAKRISHNAN, *A class of discontinuous Petrov–Galerkin methods. II. Optimal test functions*, Numer. Methods Partial Differential Equations, 27 (2011), pp. 70–105.
- [11] ———, *An overview of the discontinuous Petrov Galerkin method*, in Recent Developments in Discontinuous Galerkin Finite Element Methods for Partial Differential Equations: 2012 John H Barrett Memorial Lectures, X. Feng, O. Karakashian, and Y. Xing, eds., vol. 157 of The IMA Volumes in Mathematics and its Applications, Springer, Cham, 2014, pp. 149–180.
- [12] A. ERN AND J.-L. GUERMOND, *Theory and Practice of Finite Element Methods*, vol. 159 of Applied Mathematical Sciences, Springer-Verlag, New York, 2004.
- [13] J. GOPALAKRISHNAN, *Five lectures on DPG methods*. arXiv:1306.0557v2 [math.NA], Aug 2014.
- [14] J. GOPALAKRISHNAN AND W. QIU, *An analysis of the practical DPG method*, Math. Comp., 83 (2014), pp. 537–552.

- [15] J. L. GUERMOND, *A finite element technique for solving first-order PDEs in L^p* , SIAM J. Numer. Anal., 42 (2004), pp. 714–737.
- [16] J. T. ODEN AND L. F. DEMKOWICZ, *Applied Functional Analysis*, CRC Press, 2nd ed., 2010.
- [17] W. V. PETRYSHYN, *A characterization of strict convexity of Banach spaces and other uses of duality mappings*, J. Funct. Anal., 6 (1970), pp. 282–291.
- [18] I. SINGER, *Best Approximation in Normed Linear Spaces by Elements of Linear Subspaces*, vol. 171 of Die Grundlehren der mathematischen Wissenschaften, Springer, Berlin, 1970.
- [19] I. STAKGOLD AND M. HOLST, *Green's Functions and Boundary Value Problems*, vol. 99 of Pure and Applied Mathematics, John Wiley & Sons, Hoboken, New Jersey, 3rd ed., 2011.
- [20] A. STERN, *Banach space projections and Petrov–Galerkin estimates*, Numer. Math., 130 (2015), pp. 125–133.
- [21] E. ZEIDLER, *Nonlinear Functional Analysis and its Applications, III: Variational Methods and Optimization*, Springer-Verlag, New York, 1985.
- [22] J. ZITELLI, I. MUGA, L. DEMKOWICZ, J. GOPALAKRISHNAN, D. PARDO, AND V. M. CALO, *A class of discontinuous Petrov–Galerkin methods. Part IV: The optimal test norm and time-harmonic wave propagation in 1D*, J. Comput. Phys., 230 (2011), pp. 2406–2432.